Indian Journal of Engineering

ANALYSIS

ISSN 2319 – 7757 EISSN 2319 –7765

Gene Expression Dataset Analysis

Publication History
Received: 04 August 2015
Accepted: 21 August 2015
Published: 1 September 2015 (Special issue)

Citation
Rajeswari R, Obulakshmi O. Gene Expression Dataset Analysis. *Indian Journal of Engineering*, 2015, 12(30), 146-151

Gene Expression Dataset Analysis

R. Rajeswari
Research scholar
Department of Computer Applications
St. Peter's University, Chennai.

O.Obulakshmi
MSc Computer Science
Salem Sowdeswari College
Salem-10.

Abstract:

Gene expression data is analyzed to identify the hidden features thereby, predicting the functioning and properties of the gene. The gene database is collected from different sources whereas, This paper provides major steps on collection of gene dataset on basis of different formats, sources, reliability and acquisition of data used for clustering technique, while summarizing their features extraction of gene data. The data which is collected in this proposed method will be further used for mining, clustering, classification etc.,

Keywords: Data mining, clustering, bio infomatics, Gene expression data,

I. Introduction:

Genes are the distinct component of DNA, it carries the genetic information that is used to make all the proteins the cell needs. Where DNA is deoxyribonucleic acid found in all cells present in the body. Proteins are the structural components of cells and tissues and it perform many key functions of biological systems and the production of proteins is controlled by genes. The genetic transformation from DNA into the primary structure of a protein by two steps are transcription is that the genetic information encoded in DNA is transferred to RNA; In which information of the cell that contained inside the cell nucleus of the cell and the next step is translation follows the movement of mRNA to the cytoplasm where it interacts with structures called ribosome's to synthesize a protein. It contains a particular set of instructions that code for a specific protein. Gene expression data is usually represented as an m × n matrix, where m is the number of genes and n is the number of samples. Microarray features, or gene transcripts, are the rows of the expression matrix and are represented as vectors. Gene expression datasets are comprised of gene expression levels overtime points, also called time course data. The techniques available for measuring gene expression, including serial analysis of gene expression (SAGE), cDNA library sequencing, cDNA subtraction, multiplex quantitative RT-PCR, and gene expression microarrays.

Microarray data have been stored in public databases such as the Gene Expression Omnibus (GEO) for further analysis. Gene set referred as a collection of genes based on knowledge of genes and biological processes. Large number of gene sets databases now available can be used for gene set analyses. DNA microarrays can be used to detect that may be translated into proteins.

Micro array data to Gene Expression data:

A microarray may contain thousands of spots. Each spot contains many copies of the same DNA sequence that uniquely represents a gene from an organism. Spots are arranged in an orderly fashion into Pengroups.

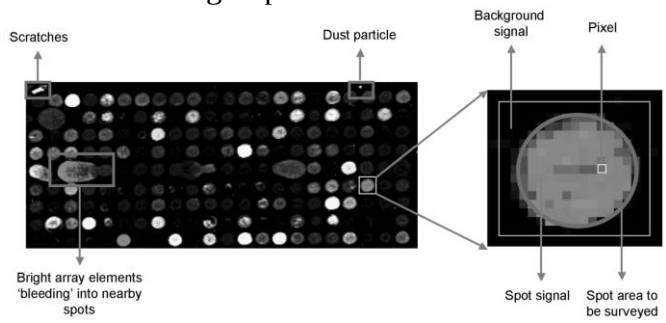


Figure - I

Information present in microarray chip is transferred as a Gene data set (Fig - I). It is further classify and cluster the gene based on the criteria. This method is widely used in many hospitals for disease identification in beginning stage. Therefore each hospital independently has their format (CLM,CHIP,POL,GMT,TXT,etc.,)to maintain their gene database. In this paper, the different format of gene expression data is analyzed.

II. Literature Survey:

Balaji Venkataraman compares the quality of the generated data with published data and is superior to that obtained using spotted oligonucleotide microarrays [1]. Erliang Zeng produces integrated heterogeneous constrained dataset [2]. Kun Gao in 2014 proposes DNA matures the gene expression data [3]. Chun-Hou Zheng in 2006 the sequential floating forward selection technique is used to select the independent components of the DNA microarray data for classification [4]. Jesse M. Engreitz in 2010 presents a unique opportunity for intelligent data mining methods to extract information about the transcriptional modules underlying these biological processes [5]. V. Sitras in 2015 share same common traits in their gene expression profile indicating common pathways in their pathophysiology [6]. Soheila Khodakarim in 2014 two new methods, in comparison to the previous ones, is introduced for GSA [7]. Kaiyang Liao, Guizhong Liu and et., produces a result on K-means algorithm is a typical partition-based clustering method [8]. Petri Toronen and et al., produce an algorithm that dealing with noisy data, and is capable of generating an appealing map of data sets in both 2D and 3D space [9] [10]. Md. Bahadur Badsha proved the clustering of gene data using hierarchical clustering algorithm [11]. Zhan C.T., and Janne Nikkila, Petri Toronen provides a graph theoretical approach treats the entire data set as a graph [12]. Adrian E Raftery and Nema Dean clusters the data set as a finite mixture of probability distributions using model-based clustering [14]. Daxin Jiang and etl., clusters the data on the basis of high-dimensional dense cluster [15].

III. Existing Model:

Gene expression data is a standard prefatory technique used for similar gene identification. Some of the data sources are also likely to be of great reinforcement in the analysis of gene expression data. Where the sources include protein interaction data, transcription factor and regulatory elements data, comparative genomics data, protein expression data and much more.

Data Sources:

Data Source is a unification tool which dynamically collects and compiles data from many scientific databases, and thereby attempts to encapsulate the genetics and molecular biology of genes from the genomes. Source is designed to facilitate the analysis of large sets of data that biologists can now produce using genome scale experimental approaches. We collected a set of data sources used in bioinformatics experiments obtained from public databases. It is aimed to find an appropriate data set to use as an input data source as a test data for the extended algorithm in order to evaluate the performance of the proposed method and the main source for gene expression data is the Affymetrix GeneChip LIMS database.

Data Format:

Microarray Gene Expression Database Group (MGED), a consortium of academic and commercial organizations with the shared goal of defining standard formats that would allow gene expression data repositories to share and exchange data. Effective exploration of microarray data has been hindered by the variety and heterogeneity of the data formats used are MAGE-ML, MAGE-OM for microarray experiment data.

Many new tools are developed and maintained which convert data from sources into a well-defined data format, such as one based on XML or similar notation, is generally easier than developing tools to transform data and populate a target data warehouse. After loading the dataset, program gives an information messages about the loading status and loaded data input file format was created for the algorithm. Each of the following data rows starts with a gene ID and is followed by a gene expression or gene location data. Gene expression data is normalized to the 0-1 range (i.e. the data needs to be normalized in order to have minimum expression value equals to 0 and the maximum expressions value equals to 1, also known as Max-Min Normalization) before it is used by the algorithm. Here the gene data is normalized by classifying the gene data set either 0 or in 1. In the Fig-II GeneID is given row and the Gene parameters are given in column.

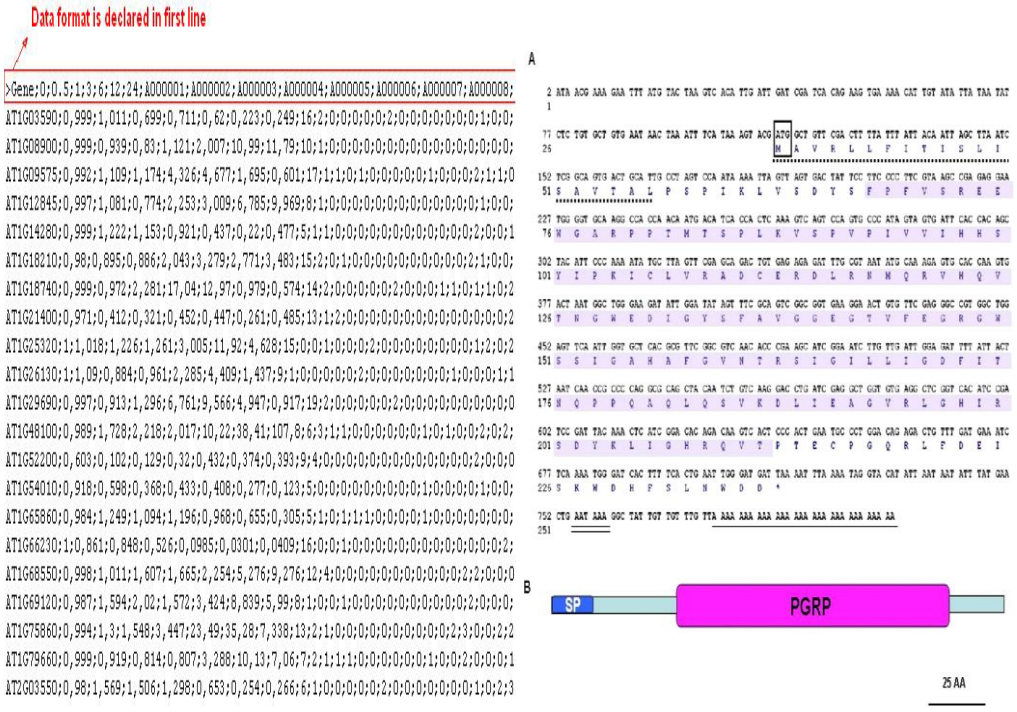


Figure – II

Data Acquisition:

Data acquisition is the process of sampling the real world signals that can be measured and convert sampling result into digital numeric values that can extracted as an input file for any classification and clustering process. Data acquisition technologies depend on parallelization rather than on reducing the time needed to take individual data points. These technologies are capable of carrying out global (or nearly global) analyses, and as such they are well suited for the rapid and comprehensive assessment of biological system properties and dynamics. Some of the existing techniques,

DNA Microarrays: This technology enables the simultaneous interrogations of a human genomic sample for complete human transcriptomes, provided that the arrays do not contain only putative protein coding regions. Automated DNA sequences. Prior to automated sequencing, the sequencing of DNA was performed manually.

Mass spectroscopy: It enables the in-quantity identification and quantification of large numbers of proteins. Used in conjunction with genomic information, MS information can be used to identify and type single-nucleotide polymorphisms.

Cell sorters: The sorters can separate different cell types at high speed on the basis of multiple parameters. While microarray experiments provide information on average levels of mRNA or protein within a cell population, the reality is that these levels vary from cell to cell.

Microfluidic systems: This systems, also known as micro-TAS (total analysis system), allow the rapid and precise measurement of sample volumes of peculator size. These systems put onto a single integrated circuit all stages of chemical analysis, including sample preparation, analyze purification, microliquid handling, analyze detection, and data analysis.

Embedded networked sensor (ENS) systems: ENS systems are large-scale, distributed systems, composed of smart sensors embedded in the physical world that can provide data about the physical world at unprecedented granularity. These systems can monitor and collect large volumes of information at low cost on such diverse subjects as plankton colonies, endangered species, and soil and air contaminants.

Data Reliability:

Reliability analysis enables filtering of microarray data before estimating the expression ratios. This reliability based filtering can dramatically reduce number of false positives. Assessment of reliability of microarray data and estimation of signal thresholds using mixture modeling. It process on the raw data, not on the expression ratios that may be based on unreliable

signals. They are classify as Univariate by optimal thresholding or Bivariate analysis by decision-boundary method of microarray data as reliable or unreliable. Therefore, reliability probability to each microarray data point and filters data based on probabilities. It can analyze both cDNA and oligonucleotide microarray data which accepts single or dual channel data. While analyzing dual channel data in bivariate mode, it takes the correlation between the channels into account, leading to further accurate assessments of reliability. With several built-in data transformation options, it is adaptable to any data distribution to find the best possible normal mixture model for the data.

IV. Proposed Model:

Expression Data format:

Gene data is representation Excel Spreadsheet, where the data can either numerical or categorical or mixed by the both format of dataset. Data are arranged in bi-directional allotment as GeneID in rows and GeneSamples in columns. The empty or 'Na' in the cells of the datasheet represent that some data value is missing from the dataset. The formats of gene data are as follows,

Phenotype Data Formats:

Gene data are represented in numerical values of 0's and 1's.Input phenotype data from a plain text file, e.g. test.phen. If the phenotypic value is coded as 0 or 1, then it will be recognized as a case-control study (0 for controls and 1 for cases). Missing value should be represented by "NA".

011	0101	0.98
012	0102	-0.76
013	0103	-0.06

Input discrete covariates from a plain text file, e.g. test.covar. Each discrete covariate is recognized as a categorical factor with several levels. The levels of each factor can be represented by a single character, word or numerical number.

01	0101	F	Adult	0
02	0203	M	Adult	0
03	0305	F	Adolescent	1

Input quantitative covariates from a plain text file, e.g. test.qcovar. Each quantitative covariate is recognized as a continuous variable.

#numeric
#AFFX-BioB-5_st
206.0 31.0 252.0 -20.0 -169.0 -66.0 230.0 -23.0 67.0 173.0 -55.0 -20.0 469.0 -201.0 -117.0 -162.0 -5.0 -86.0 350.0 74.0 -215.0 193.0 506.0 183.0 350.0 113.0 -17.0 29.0 247.0 -131.0 358.0 561.0 24.0 524.0 167.0 -56.0 176.0 320.0

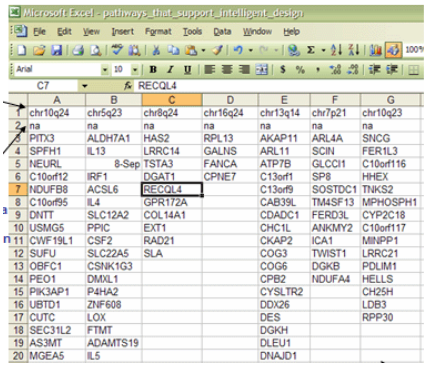
Here the list of gene values is represented in sequential or continuous values.

Gene set Database format:

Gene dataset is represented in rows and columns in excel spreadsheet. The GRP file format contains a single gene set in a simple newline-delimited text format. The GMT or GMX file formats to create gene sets, rather than using the GRP file format. The GRP format contains a line for each gene, one gene per line. Lines that start with a pound sign (#) are ignored.

P1 (reverse)	CTAACTAATTTTATTGGACTAGGC
P2 (forward)	TTCGTAAGCCGAGAGGAATGGGG
P3 (forward)	TCTCTGTGCTGTGAATAACT
P4 (reverse)	ACAAACAACAATAGCCTTT
Oligo (dG)-adaptor	GGCCACGCGTCGACTAGTACG10
Expression primers	
P5	CGCGGATCCTTGCCTAGTCCAATAAAATTAG
P6	CCGCTCGAGTTAATCATCCCAATTCAGTG
RT primers	
RTF	GTCCAGTGCCCATAGTAGTGAT
RTR	CGATGCTTCGGGTGTTG
Oligo (dT)-adaptor	GGCCACGCGTCGACTAGTACT17
Actin primers	
AF (forward)	CCGTATGCGAAAGGAAATCA
AR (reverse)	TTGGAAGGTAGAGAGGGAGG
Vector primers	
M13–21 (forward)	TGTAAAACGACGGCCAGT
M13− (reverse)	CAGGAAACAGCTATGACC

In the example the gene data is represented in text format with its GeneID simultaneously. Though there are many other gene formats like, Microchip Annotation Format, Ranked Gene List Format, XCN File Format, etc., the Text File Format is better method for further work on gene classification and Clustering to identify the 'Na' or missing values.



	A	B	C	D	E	F	G
1	chr10q24	chr5q23	chr8q24	chr16q24	chr13q14	chr7p21	chr10q23
2	na	na	na	na	na	na	na
3	PITX3	ALDH7A1	HAS2	RPL13	AKAP11	ARL4A	SNCG
4	SPFH1	IL13	LRRC14	GALNS	ARL11	SCIN	FER1L3
5	NEURL	8-Sep	TSTA3	FANCA	ATP7B	GLCCI1	C10orf116
6	C10orf12	IRF1	DGAT1	CPNE7	C13orf1	SP8	HHEX
7	NDUFB8	ACSL6	RECQL4		C13orf9	SOSTDC1	TNKS2
8	C10orf95	IL4	GPR172A		CAB39L	TM4SF13	MPHOSPH1
9	DNTT	SLC12A2	COL14A1		CDADC1	FERD3L	CYP2C18
10	USMG5	PPIC	EXT1		CHC1L	ANKMY2	C10orf117
11	CWF19L1	CSF2	RAD21		CKAP2	ICA1	MINPP1
12	SUFU	SLC22A5	SLA		COG3	TWIST1	LRRC21
13	OBFC1	CSNK1G3			COG6	DGKB	PDLIM1
14	PEO1	DMXL1			CPB2	NDUFA4	HELLS
15	PIK3AP1	P4HA2			CYSLTR2		CH25H
16	UBTD1	ZNF608			DDX26		LDB3
17	CUTC	LOX			DES		RPP30
18	SEC31L2	FTMT			DGKH		
19	AS3MT	ADAMTS19			DLEU1		
20	MGEA5	IL5			DNAJD1		

Figure - III

V. Conclusion:

In this proposed method is implemented by collecting different types of constrained data, from any hospital database or online gene database sources in text format and to produce the result by using MATLAB, Rapid miner as a supporting tool. This type of gene dataset collection is further used for research on gene data analysis and those genes

can be classified , cluster or any other data mining process.

VI. References:

[1] Balaji Venkataraman, Madavan Vasudevan, Amita Gupta "A new microarray platform for whole-genome expression profiling of Mycobacterium tuberculosis" in Journal of Microbiological Methods 97 (2014) 34–43.

[2] Erliang Zeng, Chengyong Yang, and etl., "Clustering Genes using Heterogeneous Data Sources" in 3Bioinformatics Research Group (BioRG), School of Computing and Information Sciences, Florida International University, 11200 SW 8th Street, Miami, FL 33199.

[3] Kun Gao, Xiang-yuan Deng, He-ying Qian, Guang-xing Qin, Cheng-xiang Hou, Xi-jie Guo" Cloning and expression analysis of a peptidoglycan recognition protein in silkworm related to virus infection" in journal homepage: www.elsevier.com/locate/gene , Gene 552 (2014) 24–31.

[4] Chun-Hou Zheng, De-Shuang Huanga, Li Shanga, "Feature selection in independent component subspace for microarray data classification" in Neurocomputing 69 (2006) 2407–2410

[5] Jesse M. Engreitz a, Bernie J. Daigle Jr. b, Jonathan J. Marshall a, Russ B. Altman" Independent component analysis: Mining microarray data for fundamental human gene expression modules" in Journal of Biomedical Informatics 43 (2010) 932–944.

[6] V. Sitras , C. Fenton , G. Acharya "Gene expression profile in cardiovascular disease and preeclampsia:A meta-analysis of the transcriptome based on raw data from human studies deposited in Gene Expression Omnibus" in Placenta 36 (2015) 170e178

[7] Soheila Khodakarim, Seyyed Mohammad Tabatabaei, Hamid AlaviMajd," The Multivariate Nonparametric Methods for Identifying Gene Sets with Differential Expression" in Gene 552 (2014) 18–23.

[8] Kaiyang Liao, Guizhong Liu, Li Xiao, and Chaoteng Liu "A sample-based hierarchical adaptive K-means clustering method for large-scale video retrieval", Knowledge-Based Systems, Elsevier,Vol 49, pp 123-133,2013.

[9] Petri Toronen, Mikko Kolehmainen, Garry Wong, and Eero Castren," Analysis of gene expression data using self-organizing maps", FEBS Letters, Federation of European Biochemical Societies,Vol 451,pp 142-146,1999.

[10] Janne Nikkila, Petri Toronen, Samuel Kaski, Jarkko Venna, Eero Castren, and Garry Wong," Analysis and visualization of gene expression data using Self-Organizing Maps", Neural Networks Elsevier,Vol 15, pp 953-966,2002.

[11] Md. Bahadur Badsha, Md. Nurul Haque Mollah, Nusrat Jahan, and Hiroyuki Kurata, "Robust complementary hierarchical clustering for gene expression data analysis by β-divergence", Journal of Bioscience and Bioengineering, Elsevier,Vol 116,No3, pp 397-407,2013.

[12] Zhan C.T.," Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters ", Computers, IEEE Transactions,Vol C-20, issue 1, pp 68-86, Jan 1971.

[13] Wu, Z. and Leahy, R., "An optimal graph theoretic approach to data clustering: theory and its application to image segmentation ", Pattern Analysis and Machine Intelligence, IEEE Transactions on, Vol 15 , Issue 11 ,pp 1101-1113,Aug 2002.

[14] Adrian E Raftery and Nema Dean, "Variable Selection for Model-Based Clustering", Journal of the American Statistical Association, Vol 101, Issue 473, pp 168-178,2006

[15] Daxin Jiang, Jian Pei, and Aidong Zhang," DHC: a density-based hierarchical clustering method for time series gene expression data", In Proceedings of Third IEEE Symposium on Bioinformatics and Bioengineering, pp 393–400, March 2003.